PATENT
5201-27300
03-1509

CERTIFICATE OF EXPRESS MAIL
UNDER 37 C.F.R. § 1.10

"Express Mail" mailing label no. EV403685420US
DATE OF DEPOSIT: November 18, 2003

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" Service Under 37 C.F.R. §1.10 on the date indicated above and is addressed to: Commissioner for Patents and Trademarks, **BOX PATENT APPLICATION**, Washington, D.C. 20231

Pamela Gerik
Pamela Gerik

AN IMPROVED MEMORY CELL ARCHITECTURE

By:

Subramanian Ramesh
1148 Elmsford Drive
Cupertino, CA 95014

Ruggero Castagnetti
152 Elliott Drive
Menlo Park, CA 94025

Ramnath Venkatraman
1162 Creekwood Drive
San Jose, CA 95129

Atty. Dkt. No. 5201-27300

Kevin L. Daffer/JMF
Conley, Rose & Tayon
P.O. Box 398
Austin, TX 78767-0398
Ph: (512) 476-1400

BACKGROUND OF THE INVENTION

1. Field of the Invention

This invention relates to semiconductor integrated devices and, more particularly, to semiconductor memory devices providing increased memory speed, performance and robustness within a highly compact memory cell layout..

2. Description of the Related Art

The following descriptions and examples are not admitted to be prior art by virtue of their inclusion within this section.

Generally speaking, system-on-chip (SoC) technology is the ability to place multiple function "subsystems" on a single semiconductor chip. The term "system-on-chip" may be used to describe many of today's complex ASICs, where many functions previously achieved by combining multiple chips on a board are now provided by one single chip. SoC technology provides the advantages of cutting development cycle time, while increasing product functionality, performance and quality. The various types of subsystems that may be integrated within the semiconductor chip include microprocessor and micro-controller cores, digital signal processors (DSPs), memory blocks, communications cores, sound and video cores, radio frequency (RF) cells, power management, and high-speed interfaces, among others. In this manner, system-on-chip technology can be used to provide customized products for a variety of applications, including low-power, wireless, networking, consumer and high-speed applications.

There are various types of semiconductor memory, including Read Only Memory (ROM) and Random Access Memory (RAM). ROM is typically used where instructions or data must not be modified, while RAM is used to store instructions or data which must not only be read, but modified. ROM is a form of non-volatile storage – i.e., the

information stored in ROM persists even after power is removed from the memory. On the other hand, RAM storage is generally volatile, and must remain powered-up in order to preserve its contents.

A conventional semiconductor memory device stores information digitally, in the form of bits (i.e., binary digits). The memory is typically organized as a matrix of memory cells, each of which is capable of storing one bit. The cells of the memory matrix are accessed by wordlines and bitlines. Wordlines are usually associated with the rows of the memory matrix, and bitlines with the columns. Raising a wordline activates a given row; the bitlines are then used to read from, or write to, the corresponding cells in the currently active row. Memory cells are typically capable of assuming one of two voltage states (commonly described as "on" or "off"). Information is stored in the memory by setting each cell in the appropriate logic state. For example, to store a bit having a value of 1 in a particular cell, one would set the state of that cell to "on;" similarly, a value of 0 would be stored by setting the cell to the "off" state. (Obviously, the association of "on" with 1 and "off" with 0 is arbitrary, and could be reversed.)

The two major types of semiconductor RAM, Static Random Access Memory (SRAM) and Dynamic Random Access Memory (DRAM), differ in the manner by which their cells represent the state of a bit. In an SRAM, each memory cell includes transistor-based circuitry that implements a bi-stable latch. A bi-stable latch relies on transistor gain and positive (i.e. reinforcing) feedback to guarantee that it can only assume one of two states – "on" or "off." The latch is stable in either state (hence, the term "bi-stable"). It can be induced to change from one state to the other only through the application of an external stimulus; left undisturbed, it will remain in its original state indefinitely. This is just the sort of operation required for a memory circuit, since once a bit value has been written to the memory cell, it will be retained until it is deliberately changed.

In contrast to the SRAM, the memory cells of a DRAM employ a capacitor to store the "on"/"off" voltage state representing the bit. A transistor-based buffer drives the capacitor. The buffer quickly charges or discharges the capacitor to change the state of the memory cell, and is then disconnected. Ideally, the capacitor then holds the charge placed on it by the buffer and retains the stored voltage level.

DRAMs have at least two drawbacks compared to SRAMs. The first of these is that leakage currents within the semiconductor memory are unavoidable, and act to limit the length of time the memory cell capacitors can hold their charge. Consequently, DRAMs typically require a periodic refresh cycle to restore sagging capacitor voltage levels. Otherwise, the capacitive memory cells would not maintain their contents. Secondly, changing the state of a DRAM memory cell requires charging or discharging the cell capacitor. The time required to do this depends on the amount of current the transistor-based buffer can source or sink, but generally cannot be done as quickly as a bi-stable latch can change state. Therefore, DRAMs are typically slower than SRAMs. However, DRAMs tend to offset these disadvantages by offering higher memory cell densities, since the capacitive memory cells are intrinsically smaller than the transistor-based cells of an SRAM.

As SoC technology becomes more sophisticated, greater density, speed and performance are demanded from memory devices embedded thereon. For this reason, SRAM devices – rather than DRAM devices – are typically used in applications where speed is of primary importance, such as in communication and networking SoC applications (e.g., routers, switches and other traffic control applications). The SRAM devices most commonly used for communication and networking SoC applications are single-port devices (FIG. 1) and dual-port devices (FIG. 2), and in some cases, two-port devices (not shown).

FIG. 1 is a circuit diagram of a typical single-port SRAM memory cell 100. In general, memory cell 100 includes six transistors and uses one bi-directional port for accessing the storage element. As shown in FIG. 1, memory cell 100 utilizes a minimum of five connections; one wordline (WL) for accessing the port, two bitlines (BL/BLB) for storing the data and data complement within the storage element, one power supply line (VDD) and one ground supply line (VSS) for powering the storage element and holding the data. The storage element, or bi-stable latch, of memory cell 100 may be implemented with cross-coupled P-channel load transistors ($T1_P$ and $T2_P$) and N-channel latch transistors ($T1_N$ and $T2_N$). In an alternative embodiment, however, resistive load devices may be used in place of the P-channel load transistors, as is known in the art. A pair of N-channel access transistors (T3 and T4) provide access to the storage nodes (SN/SNB) of the bi-stable latch.

In some cases, memory cell 100 may be accessed by applying a positive voltage to the wordline (often referred to as "raising the wordline"), which activates access transistors T3 and T4. This may enable one of the two bitlines (BL/BLB) to sense the contents of the memory cell based on the voltages present at the storage nodes. For example, if storage node SN is at a high voltage (e.g., a power supply voltage, VDD) and node SNB is at a low voltage (e.g., a ground potential, VSS) when the wordline is raised, latch transistor $T2_N$ and access transistor T4 are activated to pull the bitline complement (BLB) down toward the ground potential. At the same time, the bitline (BL) is pulled up towards the power supply voltage by activation of latch transistor $T1_P$ and access transistor T3. In this manner, the state of the memory cell (either a "1" or "0") can be determined (or "read") by sensing the potential difference between bitlines BL and BLB. Conversely, writing a "1" or "0" into the memory cell can be accomplished by forcing the bitline or bitline complement to either VDD or VSS and then raising the wordline. The potentials placed on the pair of bitlines will be transferred to respective storage nodes, thereby forcing the cell into either a "1" or "0" state.

Some SoC applications benefit from the use of dual-port or two-port memory devices, which allow two independent devices (e.g., a processor and micro-controller, or two different processors) to have simultaneous read and/or write access to memory cells within the same row or column. Dual-port and two-port memory devices are essentially identical in form, and as such, can both be described in reference to FIG. 2. However, dual-port and two-port memory devices differ in function. Where both ports are used for read and write operations in a dual-port cell, one port of a two-port cell is strictly used for a write operation, while the other port of a two-port cell is strictly used for a read operation.

FIG. 2 is a circuit diagram of a typical dual-port SRAM memory cell 200, which utilizes a pair of bi-directional ports – referred to as "port A" and "port B" – for accessing the storage element. As shown in FIG. 2, memory cell 200 utilizes eight connections, including one wordline (WL_A/WL_B) for accessing each port and two pairs of bitlines (BL_A/BLB_A and BL_B/BLB_B) for reading/writing to the nodes of the storage element, as well as, a power supply line (VDD) and ground supply line (VSS). In addition to the six transistors described above for single-port memory cell 100, a second pair of N-channel access transistors (T5 and T6) are included within dual-port memory cell 200 for accessing storage nodes SN and SNB via the additional port.

Like most semiconductor devices, SRAM devices are typically fabricated en masse on semiconductor wafers over numerous processing steps. For example, an SRAM device may be fabricated as a metal-oxide-semiconductor (MOS) integrated circuit, in which a gate dielectric, typically formed from silicon dioxide (or "oxide"), is formed on a semiconductor substrate that is doped with either n-type or p-type impurities. Conductive regions and layers of the device may also be isolated from one another by an interlevel dielectric. For each MOS field effect transistor (MOSFET) within the SRAM device, a gate conductor is formed over the gate dielectric, and dopant impurities are introduced into the substrate to form "source" and "drain" regions. Frequently, the integrated circuit will employ a conductive layer to provide a local interconnect function between the

transistors and other components of the device, such as overlying bitlines, wordlines, power and ground supply lines.

A pervasive trend in modern integrated circuit manufacture is to produce transistors that are as fast as possible, and thus, have feature sizes as small as possible. Many modern day processes employ features, such as gate conductors and interconnects, which have less than 1.0 μm critical dimension. As feature sizes decrease, sizes of the resulting transistor and interconnects between transistors decrease. Fabrication of smaller transistors may allow more transistors to be placed on a single monolithic substrate, thereby allowing relatively large circuit systems to be incorporated onto a single, relatively small semiconductor chip.

As transistor feature sizes continue to decrease with advancements in manufacturing processes, greater amounts of memory may be incorporated onto the chip without increasing the chip area. This may be especially advantageous in many SoC applications, where the demand for on-chip memory is expected to increase from about 50% to about 90% of the total chip area. In an effort to effectively utilize chip area, many SoC designs divide the memory device into numerous memory blocks, which are then embedded at various locations within the chip, rather than concentrated in one large memory unit. Unfortunately, many of these SoC designs suffer from data corruption, which may be caused by stray capacitances from chip-level signals routed over the memory blocks. Though strict routing restrictions may be imposed to avoid data corruption, such restrictions often lead to chip-level routing congestion and undesirable increases in overall chip area. Therefore, a need exists for an improved memory cell architecture, which significantly decreases memory device area and chip-level routing congestion, while maintaining performance and speed specifications for next-generation SoC applications.

SUMMARY OF THE INVENTION

The problems outlined above may be in large part addressed by an improved memory cell architecture providing increased memory speed, performance and robustness within a highly compact memory cell layout. Though only a few embodiments are provided herein, a feature common to all embodiments includes a novel means for sharing one or more contact structures between vertically adjacent memory cells.

In one embodiment, a memory array includes a plurality of memory cells arranged in one or more rows and columns. In general, each memory cell within the array shares at least one contact structure with a vertically adjacent memory cell. Such a contact structure may be referred to herein as a "shared contact structure," and is generally formed proximate to a boundary (or "cell pattern boundary") between the memory cell and the vertically adjacent memory cell.

In a preferred embodiment, the contact structure may be shared unequally between the memory cell and the vertically adjacent memory cell. In some cases, for example, the shared contact structure may be: i) formed completely within the memory cell on one side of the boundary; ii) formed completely within the second memory cell on an opposite side of the boundary; or iii) formed at the boundary, such that unequal portions of the shared contact structure are formed on either side of the boundary. For example, the shared contact structure may be a bitline contact structure for coupling an overlying bit line to an underlying diffusion region, a ground supply contact structure for coupling an overlying ground supply line to an underlying diffusion region, and/or a power supply contact structure for coupling an overlying power supply line to an underlying diffusion region. By sharing the one or more contact structures unequally between the memory cell and the vertically adjacent memory cell, the present embodiment aids in reducing memory cell density by reducing a length of the memory array (e.g., by approximately 10% to 20%).

In order to accommodate the shared contact structures, a mirroring technique may be applied to incorporate the memory cell architecture into a memory array layout. In some cases, a column of memory cells may be formed by rotating vertically adjacent memory cells about an x-axis and about a y-axis, wherein the x- and y-axes extend horizontally and vertically, respectively, through a center of each memory cell. In this manner, multiple rows of memory cells may be formed by replicating the column of memory cells at a location horizontally adjacent to the column.

In a preferred embodiment, the column of memory cells may include a pair of n-type diffusion regions, where each n-type diffusion region is formed as a continuous line of constant width and periodically interspersed with rectangular shaped isolation regions. Thus, the present embodiment aids in reducing a complexity of the memory array by avoiding diffusion regions formed with complex geometries.

In a preferred embodiment, each memory cell in the column may include a first local word line and a second local word line, where each may extend only partially across each memory cell. More specifically, each memory cell may include a second local word line, which is split into distinct portions and arranged on opposite sides of the memory cell. In addition, each memory cell may include two access transistors, which share the first local word line, and an additional access transistor, which shares a portion of the second local word line with an access transistor in a horizontally adjacent memory cell. In order to isolate transistors within a common cell, a distal end of the first local word line may be horizontally and vertically spaced from a distal end of a portion of the second local word line over one of the rectangular shaped isolation regions. In some cases, an isolation region may also be shared between two vertically adjacent memory cells. Thus, the present embodiment further aids in reducing memory cell density by sharing portions of the second local word line with adjacent memory cells (i.e., reducing the width), and by terminating two or more local wordlines in a staggered formation above the isolation regions of the memory array (i.e., reducing the width and length).

As a result, one or more of the above-mentioned embodiments may be used to provide a memory cell aspect ratio of substantially less than 1.0, which may be desirable for achieving higher performance, high-density memory devices. In a preferred embodiment, the above-mentioned embodiments may be combined to provide an SRAM memory cell architecture with an aspect ratio between about 0.3 and about 0.7.

In another embodiment, a dual-port memory cell may include a first pair of N-channel access transistors coupled through respective gate terminals by a first local word line of the memory cell, and a second pair of N-channel access transistors coupled through respective gate terminals by separate portions of a second local word line of the memory cell. The dual-port memory cell may also include a plurality of bitline contact structures coupled to drain terminals of the first and second pairs of access transistors and to drain terminals of corresponding pairs of access transistors arranged within a vertically adjacent memory cell. In a preferred embodiment, the bitline contact structures may be formed i) completely within the memory cell; ii) completely within the adjacent memory cell; or iii) having unequal portions within the memory cell and the adjacent memory cell.

In general, the first and second pairs of access transistors may be coupled for accessing a storage element of the dual-port memory cell. In some cases, the storage element may include first and second inverter circuits, where each inverter circuit includes a P-channel latch transistor coupled in common-gate configuration with an N-channel latch transistor. More specifically, the drain terminals of the P-channel and N-channel latch transistors may be coupled to respective source terminals of the first and second pairs of N-channel access transistors. In addition to the bitline contact structures described above, the dual-port memory cell may also include a pair of power supply contact structures and a pair of ground supply contact structures. The pair of power supply contact structures may be coupled to source terminals of the P-channel latch transistors, while the pair of ground supply contact structures may be coupled to source terminals of the N-channel latch transistors. In a preferred embodiment, the pairs of power supply and ground supply contact structures may be shared unequally between

vertically adjacent memory cells. For example, one power supply contact structure and one ground supply contact structure may be arranged within each of the vertically adjacent memory cells.

The dual-port memory cell may further include a first metal layer, which is dielectrically spaced above and coupled to the access transistors and the latch transistors through corresponding contact structures. Such a first metal layer may also be referred to herein as a "local interconnect layer." Next, a second metal layer may be dielectrically spaced above and coupled to the first metal layer through a plurality of vias. In some cases, the second metal layer may include a first pair of complementary bit lines directed along a length of the memory cell and corresponding to a first port, and a second pair of complementary bit lines directed along the length of the memory cell and corresponding to a second port. In other words, all bitlines of the dual-port memory cell are formed within the second metal layer. The second metal layer may also include a pair of ground supply lines, each directed along the length of the memory cell and arranged between bit lines of dissimilar ports. Thus, the present embodiment may reduce intrinsic as well as cross-coupling bitline capacitances. The intrinsic bitline capacitance may be reduced by forming the bitlines in the lowest available metallization layer of the memory array (e.g., the second metal layer of an SRAM cell), whereas the cross-coupling bitline capacitances may be reduced by providing horizontal capacitive shielding (i.e., ground supply lines) between bitlines of dissimilar ports.

The dual-port memory cell may further include a third metal layer, which is dielectrically spaced above and coupled to the second metal layer through another plurality of vias. In some cases, the third metal layer may include a first word line directed along a width of the memory cell and corresponding to the first port, and a second word line directed along the width of the memory cell and corresponding to the second port. Thus, the present embodiment may further reduce cross-coupling bitline capacitances by forming wordlines within an inter-level metallization layer of the memory array (e.g., the third memory layer of an SRAM cell). In other words, the

wordlines may be used to vertically shield the bitlines from cross-coupling capacitances within an upper-level metallization layer (e.g., a fourth metal layer dielectrically spaced above and coupled to the third metal layer).

In a preferred embodiment, the third metal layer may also include a ground supply line directed along the width of the memory cell. In this manner, the ground supply line may also function to vertically shield the bitlines from cross-coupling capacitances within upper-level metallization layers. In a preferred embodiment, the ground supply line may be arranged between the first and second word lines to reduce cross-coupling wordline capacitances by providing horizontal capacitive shielding between wordlines of dissimilar ports. In some cases, the ground supply line within the third metal layer may be coupled to the pair of ground supply lines within the second metal layer to form a two-dimensional ground supply grid. As a result, a robustness of the memory cell may be increased. In some cases, the third metal layer may further include a power supply line. The power supply line may be directed along the width of the memory cell, such that portions of the power supply line are shared between the memory cell and the adjacent memory cell. Thus, the present embodiment may aid in reducing memory cell density by further decreasing the length of the memory cell. The power supply line may also function to vertically shield the bitlines from cross-coupling capacitances within upper-level metallization layers.

In yet another embodiment, a system embedded within and/or arranged upon a single substrate may include a memory array comprising a plurality of memory cells. The plurality of memory cells may be formed as described above. For example, each of the memory cells may include a substrate layer and at least two metal layers arranged above the substrate layer, though three metal layers may be preferred in some embodiments. The substrate layer may include, e.g., four access transistors, two inverter circuits, and a plurality of contact structures. As noted above, one or more of the plurality of contact structures may be shared between vertically adjacent memory cells. These shared contact

structures are preferably formed offset from a boundary (i.e., a "cell pattern boundary") extending between the vertically adjacent memory cells.

In some cases, the second metal layer may be arranged above the substrate layer. If used to form an SRAM array, for example, the second metal layer may be coupled to the substrate layer through a first metal layer or "local interconnect layer." However, it is worth noting that the local interconnect layer may not be necessary in other cases (e.g., within other types of memory arrays, such as magnetic RAMs, MRAMs). In any case, the second metal layer may include a plurality of bitlines and one or more ground supply lines. The ground supply lines are preferably arranged between and parallel to bitlines of dissimilar ports.

In some cases, the third metal layer may be arranged above the second metal layer. The third metal layer preferably includes a pair of wordlines and an additional ground supply line arranged between and parallel to wordlines of dissimilar ports. In such a case, the additional ground supply line may be parallel to the wordlines and perpendicular to the plurality of bit lines. The system may further include one or more subsystems coupled to the memory array through a fourth metal layer arranged above the third metal layer. In some cases, the fourth metal layer comprises a plurality of transmission lines. In this manner, the third metal layer may be configured to vertically shield the second metal layer from stray capacitances from the fourth metal layer.

BRIEF DESCRIPTION OF THE DRAWINGS

Other objects and advantages of the invention will become apparent upon reading the following detailed description and upon reference to the accompanying drawings in which:

FIG. 1 is a circuit schematic diagram of a single-port SRAM memory cell;

FIG. 2 is a circuit schematic diagram of a dual-port SRAM memory cell;

FIG. 3 is a top view of one embodiment of a single-port SRAM memory cell;

FIG. 4 is a top view of another embodiment of a single-port SRAM memory cell;

FIG. 5 is a cross-sectional view through line AA of FIG. 3;

FIG. 6 is a top view of one embodiment of a single-port SRAM memory cell in accordance with the present invention;

FIG. 7 is a cross-sectional view through line BB of FIG. 6;

FIG. 8 is a top view of another embodiment of a single-port SRAM memory cell in accordance with the present invention;

FIG. 9 is a cross-sectional view through line CC of FIG. 8;

FIG. 10 is a top view of an exemplary system having various subsystems and memory blocks interconnected on an upper-level metallization layer of the system;

FIG. 11 is a top view of various semiconductor layers (e.g., substrate through first metal layers) within a portion of a dual-port SRAM memory array;

FIG. 11A is a magnified view of the memory cell located within row R_{X+1}, column C_X of the portion shown in FIG. 11;

FIG. 12 is a top view of various semiconductor layers (e.g., first metal through second metal layers) within a portion of a dual-port SRAM memory array; and

FIG. 13 is a top view of various semiconductor layers (e.g., second metal through third metal layers) within a portion of a dual-port SRAM memory array.

While the invention is susceptible to various modifications and alternative forms, specific embodiments thereof are shown by way of example in the drawings and will herein be described in detail. It should be understood, however, that the drawings and detailed description thereto are not intended to limit the invention to the particular form disclosed, but on the contrary, the intention is to cover all modifications, equivalents and alternatives falling within the spirit and scope of the present invention as defined by the appended claims.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

FIGS. 3-5 illustrate exemplary embodiments of a cell architecture that may be used to form a single-port SRAM cell, such as memory cell 100 of FIG. 1. FIGS. 3 and 4 present top-down views of cell architectures 300 and 400, respectively. FIG. 5 presents a cross-sectional view along line AA of cell architecture 300. Though the architecture of only one memory cell is illustrated in FIGS. 3-5, and later in FIGS. 6-9, one of ordinary skill in the art would understand how the architecture could be applied to an array of memory cells.

In most SRAM architectures, including those illustrated in FIGS. 3-5, the bitlines and wordlines of the memory cell run orthogonal to one another on separate metallization layers. Therefore, at least two metallization layers are needed to form an SRAM cell. In some cases, more than two metallization layers may be used in an effort to increase cell density by decreasing the cell aspect ratio. In general, the term "aspect ratio" is used herein to describe the ratio between the length and the width (denoted as L:W or W/L) of a semiconductor feature. In the case of a "cell aspect ratio," the ratio is taken between the length and width of the memory cell, where "length" is defined along the bitline direction and "width" is defined along the wordline direction. An aspect ratio of less than 1.0 is often desired to achieve higher performance, high-density SRAM architectures. By reducing the length of the memory cell (i.e., forming shorter bitlines than wordlines), the RC delay through the bitlines can be reduced to provide significantly faster memory cell addressing times than can be achieved with a memory cell having an aspect ratio of 1.0 or greater.

In some cases, three metallization layers (denoted M_{X-1}, M_X, and M_{X+1}) may be incorporated into the memory cell architecture, as shown in FIGS. 3-5. In some cases, the wordlines (WL) of a memory array may be formed within what is sometimes referred to as a "first metal layer." However, since the "first metal layer" typically refers to the first conductive layer above an underlying local interconnect layer, it may be a misnomer in those cases where the local interconnect layer is formed from a metallic material. To avoid confusion, the current discussion refers to the "first metal layer" as the conductive layer, which is dielectrically spaced above the various layers and/or structures forming the storage element of the memory cell (e.g., the PMOS and NMOS transistors in an SRAM cell). As used herein, the term "dielectrically spaced" refers to the formation of an interlevel dielectric layer between two conductive layers, so as to electrically isolate the two layers.

A local interconnect layer is formed within the first metal layer of FIGS. 3-5. As a result, the wordlines of FIGS. 3-5 are formed within a "second metal layer" of the memory array, where the second metal layer is dielectrically spaced above the first metal layer. In some cases, the power supply lines (VDD) of the memory array are formed parallel to, and on the same metallization layer as the wordlines. Such a case is illustrated in FIGS. 3-5. However, the VDD lines may be alternatively formed in any direction, and within any available metallization layer of the memory array.

As noted above, the bitlines (BL/BLB) of a memory array are typically formed perpendicular to, and on a different metallization layer than the wordlines. In the embodiment of FIGS. 3-5, the bitlines are formed within a "third metal layer" of the memory array, where the third metal layer is dielectrically spaced above the second metal layer. In most cases, the ground supply lines (VSS) of a memory array are formed parallel to, and on the same metallization layer as the bitlines. In some cases, one VSS line may be formed between the BL and BLB of each column of memory cells, as shown in FIGS. 3 and 5. In other cases, one VSS line may be shared between two adjacent columns of memory cells, as shown in FIG. 4. In any case, the VSS lines are typically formed in the bitline (i.e., column) direction so as to minimize the amount of current discharged to a particular VSS line during a read operation. During operation of an SRAM array, for example, all bitlines may be precharged to a logical high state until a read or write operation is conducted. When a specific address (i.e., row and column) is asserted for a read operation, all bitlines of the memory array are discharged through respective cells coupled to the asserted wordline. By forming VSS lines parallel to bitlines, as is typically the case, each VSS line carries the current of only one cell (or two cells, in the case of FIG. 4). Therefore, this approach is generally used to reduce the occurrence of voltage droop, ground bounce and electromigration effects within the memory array.

On the other hand, if the VSS lines were formed perpendicular to the bitline direction, the current from all bitlines would be discharged onto the VSS line associated with the asserted wordline. Clearly, this configuration could lead to a potentially large amount of current discharged onto a single VSS line. Assume, e.g., that an SRAM array contains 1024 columns. If VSS lines are formed perpendicular to these columns, the current of 1024 cells would be discharged onto a single VSS line. Assuming a cell current of, e.g., 50 μA, approximately 51.2 mA of current would be discharged onto the same VSS line. Thus, a relatively wide VSS line may be required to prevent voltage droop, ground bounce, or electromigration problems in the embodiment of FIGS. 3-5.

As used herein, the term "voltage droop" refers to a voltage level on a power supply line, which drops below the level applied at a certain pin as a result of the amount of current present on the power supply line and the finite resistance of that line (i.e., Ohm's Law: V = R*I). The term "ground bounce" is used when the ground plane (usually the silicon substrate) exhibits a localized voltage that is higher than the ground potential. Ground bounce can be triggered when relatively high currents are injected into the substrate. In other words, the ground potential may "float up" when the high currents injected into the substrate are not effectively sourced to ground (e.g., due to a relatively resistive ground connection). The term "electromigration" refers to the transport of material with electrical current, and in some cases, may result in a missing metal defect (i.e., open circuit) or extra metal defect (i.e., short circuit) in a conductive line. In order to avoid electromigration, guidelines are generally used to determine the maximum amount of current allowed through a conductive line, via or contact.

Unfortunately, the memory cell architectures described above and illustrated in FIGS. 3-5 suffer from many disadvantages. As features within memory cell architectures 300 and 400 are reduced (e.g., using 130 nm technology and below), chip designers are forced to move the bitlines and VSS lines to the highest metallization layer provided by the memory cell (e.g., the third metal layer, as shown in FIGS. 3-5) in order to maintain the advantages of running the VSS lines parallel to the bitlines. In some cases, additional

layers of metal may be added to the memory cell to achieve necessary reductions in cell size. However, since bitline capacitance tends to increase at higher metallization layers, these methods often result in longer RC delays and slower memory cell addressing times (sometimes as much as 10% to 30% slower).

As another disadvantage, the above approach may require significant routing restrictions to be imposed on the next higher metallization layer (e.g., a fourth metal layer) to avoid data corruption during a read or write operation. In some cases, for example, the next higher metallization layer may be used for chip-level signal and power routing in a System-on-Chip (SoC) environment. If routing restrictions are not imposed in such an environment, capacitive coupling between the bitlines and overlying transmission lines (which may route chip-level signals at higher clock speeds) could disturb signal development on the bitlines, thereby corrupting data "sensed" during a read operation. For this reason, routing restrictions are usually imposed to prevent transmission lines from being routed above bitlines, as shown in FIG. 5. However, routing restrictions often lead to an undesirable increase in overall chip area (e.g., 15% to 20%), which is counterproductive to the desire for minimum cell size and maximum cell density.

Therefore, a need exists for an improved memory cell architecture that alleviates routing congestion within upper-level metallization layers (e.g., chip-level routing layers) of an SoC device, while allowing feature sizes within the memory cell to be aggressively scaled for reducing cell size and increasing cell density. Although the improvements described herein may be applied to stand-alone memory devices, the improvements are particularly applicable to SoC memory devices, due to the unique requirements (e.g., subsystem placement, timing, etc.) placed on each design.

FIGS. 6-9 illustrate exemplary embodiments of an improved memory cell architecture in accordance with the present invention. In particular, FIG. 6 presents a top-down view of cell architecture 500, while FIG. 7 presents a cross-sectional view along

line BB of FIG. 6. Likewise, FIG. 8 presents a top-down view of cell architecture 600, while FIG. 9 presents a cross-sectional view along line CC of FIG. 8. Though only a few embodiments are provided for the sake of brevity, features common to the embodiments described herein include: the formation of bitlines in lower-level metallization layers, and the use of ground supply lines (VSS) for effective shielding of the bitlines against routing signals in upper-level metallization layers.

FIGS. 6-9 are used herein to describe various features of the present invention in the context of a single-port CMOS SRAM cell architecture. However, the improvements described herein are not limited to such an architecture, and may be additionally applied to: SRAM cell architectures having more than one port, SRAM cell architectures formed according to different process technologies (e.g., Silicon On Insulator, SOI), other semiconductor memory cell architectures (e.g., DRAM and various non-volatile memories, such as Ferroelectric RAM, FeRAM, or Magnetic RAM, MRAM), and other semiconductor devices (e.g., analog or mixed signal elements, and CMOS based sensor elements, such as temperature, pressure, magnetic and chemical sensors). Additional features of the present invention, which may be easier to describe in the context of a memory array – rather than a single memory cell – will be provided in the discussion of FIGS. 11-13.

For the sake of clarity, FIGS. 6-9 illustrate only the metallization layers of the memory cell. However, one of ordinary skill in the art would understand that the illustrated layers are formed above various underlying semiconductor layers and structures, such as, e.g., active regions, isolation regions, polysilicon structures, and contact structures, which may be used to form the NMOS and PMOS transistors of a typical CMOS SRAM. A preferred layout of the various underlying semiconductor layers and structures will be described in more detail in FIG. 11.

As noted above, at least two metallization layers (e.g., M_X and M_{X+1}) are needed to form an SRAM cell. In one embodiment, the bitlines (BL/BLB) of the memory array may be formed within a "first metal layer." As used herein, the "first metal layer" refers to the first conductive layer, which is dielectrically spaced above the various layers and/or structures forming the storage element of the memory cell (e.g., the PMOS and NMOS transistors of an SRAM cell). As a result, the wordlines (WL) of the memory array may be formed within a "second metal layer" of the memory array, where the second metal layer is dielectrically spaced above the first metal layer. As noted above, the term "dielectrically spaced" refers to the formation of an interlevel dielectric layer between two conductive layers, so as to electrically isolate the two layers. This embodiment may be especially applicable to other types of memory cells, such as DRAM and MRAM cells, or larger memory cells.

In some cases, one or more local interconnects may be formed within the first metal layer, or alternatively, within an underlying process layer. Local interconnects are often used for short connections between conductive lines, as compared to the much longer conductive lines used for global connections (such as, e.g., bitlines, wordlines, power and ground supply lines). For example, local interconnects may be used for cross-coupling internal nodes of the NMOS and PMOS transistors used to form the SRAM cell. However, the term "local interconnect" may have multiple meanings.

In some cases, the term "local interconnect" may refer to the function of connecting features within a circuit. Such a definition may be used to describe a local interconnect formed within an underlying process layer, which is not considered a "local interconnect layer" even though the process layer may perform local interconnecting functions. In other cases, the term "local interconnect" may refer to a distinct process layer, i.e., a local interconnect layer, which exclusively performs short connections between conductive lines. Forming a distinct local interconnect layer may be desired in embodiments, which strive to conserve or reduce chip-level metal layers, and is commonly used in 0.25 μm process technologies and below. Regardless, the term "local"

may be used herein to reference a connection that extends only partially across a memory cell, whereas the term "global" refers to a connection that extends across multiple memory cells (e.g., a block of memory cells, or an entire memory array).

In other embodiments, the bitlines of the memory array may be formed within the "second metal layer" when one or more local interconnects within the "first metal layer" form a distinct "local interconnect layer". The wordlines (WL) of the memory array may then be formed within a "third metal layer" of the memory array, where the third metal layer is dielectrically spaced above the second metal layer. Such an embodiment may be utilized in high-density SRAM arrays, due to the potential difficulty in forming bitlines and cross-coupling local interconnects within the same metal layer. On the other hand, a distinct local interconnect layer may not be needed to form other types of memory devices, such as DRAMs and MRAMs, or larger-sized memory devices (using, e.g., 0.3 μm process technologies and above).

Thus, one feature of the present invention is the formation of all bitlines within the lowest metallization layer appropriate for a particular type of memory and process technology. By forming all bitlines within the second metal layer (or lower), the intrinsic capacitance of the bitlines can be reduced to attain faster memory addressing times. By forming wordlines above the bitlines, chip-level signals (CLS) can be routed over the memory device without the risk of disturbing signal development on the bitlines during a read operation. Such an advantage will be described in more detail below.

Regardless of the layer on which they reside, the bitlines are preferably arranged along a first direction (e.g., the column direction), while the wordlines are arranged along a second direction (e.g., the row direction) of the memory array. In most cases, the second direction is substantially perpendicular to the first direction, where "substantially perpendicular" is described as an angular difference in the vicinity of 90°. However, the angular difference between the two directions may be slightly less, or slightly greater,

than 90°. For example, the angular difference between the two directions may range between about 45° and about 135° (especially in the case of magnetic memory cells).

As noted above, the power supply lines (VDD) of a memory array may be formed along any direction, and within any available metallization layer of the memory array. In the embodiments of FIGS. 6-9, the VDD lines are formed parallel to the wordlines (i.e., in the second direction), and on the same metallization layer as the wordlines (i.e., within layer M_{X+1}). By forming the VDD lines above the bitlines, the VDD lines may also protect signal development on the bitlines during a read operation. In some cases, the VDD lines may also provide capacitive shielding between wordlines of dissimilar port, as will be described in more detail below.

In a preferred embodiment, at least a portion of the ground supply lines (VSS_1) are arranged along the second direction of the memory array within a metallization layer, which resides above the bitline metallization layer (i.e., layer M_X). In other words, at least a portion of the ground supply lines are formed substantially perpendicular to, and on a different metallization layer than the bitlines. In the embodiments of FIGS. 6-9, the VSS_1 lines are formed on the same metallization layer as the wordlines (i.e., layer M_{X+1}). In alternative embodiment, the VSS_1 lines may be formed on a different metallization layer than the wordlines (e.g., layer M_{X+2}). However, a minimum number of metallization layers is generally preferred to reduce manufacturing costs (by reducing the number of processing steps) and increase the speed of the memory device.

Thus, another feature of the present invention is the use of wordlines and ground supply lines as effective shielding of bitlines against signals routed above and/or across the memory device. For example, one or more transmission lines used for chip-level signal and power routing may be formed within an upper-level metallization layer in a System-on-Chip (SoC) environment. To ensure proper functioning of the memory array during a read operation, the wordlines and at least a portion of the ground supply lines (VSS_1) are formed within one or more inter-level metallization layers, i.e., one or more

metal layers arranged between the bitline (lower-level) and the transmission line (upper-level) metallization layers. In this manner, the wordlines and VSS_1 lines provide vertical shielding between the bitlines and transmission lines, and thus, function to substantially eliminate cross-coupling capacitances therebetween. By protecting bitline signal development during read operations, the vertical shielding provided by the wordlines and VSS_1 lines reduces the occurrence of data corruption in the "sensed" signal. As a result, undesirable increases in chip area are avoided, since strict routing restrictions are no longer needed to ensure proper memory operation.

Because the VSS_1 lines are perpendicular to the bitlines of the memory array, a substantially large amount of current may be discharged onto a single VSS_1 line during a read operation. To accommodate this potentially large discharge current, another portion of ground supply lines (VSS_2) are arranged along the first direction of the memory array. In other words, another portion of the ground supply lines (VSS_2) may be formed substantially parallel to the bitlines of the memory array. In doing so, the adverse effects of voltage droop, ground bounce and electromigration can be reduced, or even avoided, by interconnecting one or more of the VSS_1 lines and VSS_2 lines to form a two-dimensional ground supply grid. Such a grid may be designed to the specifications of any memory array by inserting the VSS_2 lines at a particular frequency, as described in more detail below. Since the VSS_1 lines and VSS_2 lines are formed on different metallization layers, they may be interconnected at periodic intervals (e.g., every 8, 16, 32... columns) through vias, which extend through the dielectric layer separating the corresponding metal layers.

In some cases, the VSS_2 lines may be formed within the same metallization layer as the bitlines (i.e., layer M_X), as shown in FIGS. 6 and 7. In some embodiments, N-number of VSS_2 lines may be inserted within every column of an N-port memory device. If the memory device includes more than one port, the VSS_2 lines are preferably arranged between bitlines of differing ports (e.g., between bitlines of Port A and bitlines of Port B) to provide horizontal capacitive shielding therebetween. Though it may be feasible to

insert a greater or lesser number of VSS_2 lines within each column, it is generally preferred that the insertion frequency and/or width of the VSS_2 lines be chosen so as to reduce the above-mentioned effects, while avoiding unnecessary increases in memory area. For example, a VSS_2 lines may be inserted between multiple columns (e.g., every 4 to 64 columns, dependent on voltage restrictions) if the memory device includes only one port (i.e., a single-port memory device).

The configuration of FIGS. 6-7 may be preferred when routing congestion and/or memory speed, rather than memory device area, are of utmost concern. In other words, routing congestion within the upper-level metallization layers may be substantially eliminated by forming the VSS_2 lines within a lower-level metallization layer of the memory array. As a result, routing restrictions for chip-level signals in upper-level metallization layers may be altogether avoided. However, memory density may be sacrificed, to some extent, since formation of the VSS_2 lines within the memory array may increase the overall area consumed by the memory array. Thus, the configuration of FIGS. 6-7 produces a relatively "wide" and "short" memory cell architecture, which may be used to minimize the number of metallization layers in a memory cell (i.e., the height of the memory cell) for maximum memory speed and performance.

In other cases, the VSS_2 lines may be formed within a different metallization layer than the bitlines, as shown in FIGS. 8 and 9. In some embodiments, the VSS_2 lines may be formed within an upper-level metallization layer (e.g., layer M_{X+2}), which is dielectrically spaced above the memory array. For example, the upper-level metallization layer may include a plurality of transmission lines for routing chip-level signals (CLS) between various blocks of memory and one or more subsystems in a System-on-Chip environment. Though the VSS_2 lines may be functionally coupled to the memory array, they are not included within metal layers of the memory array in the configuration of FIGS. 8-9.

The configuration of FIGS. 8-9 may be preferred when memory device area, rather than routing congestion and memory speed, is of utmost concern. In other words, the amount of area consumed by the memory array may be significantly reduced by forming the VSS_2 lines above the metal layers of the memory array. Although formation of the VSS_2 lines within the upper-level metallization layer may somewhat limit routing through that layer, routing channels between the VSS_2 lines will still be available for mitigating chip-level routing congestion (e.g., by tailoring chip-level signal and power routing to comply with the requirements of the memory array). Thus, the configuration of FIGS. 8-9 produces a relatively "narrow" and "tall" memory cell architecture, which may be used to maximize memory cell density at the cost of lower memory speed and performance.

In the configuration of FIGS. 8-9, the appropriate insertion frequency and/or width of the VSS_2 lines can be chosen for substantially eliminating voltage droop, ground bounce, and electromigration effects without affecting memory density. For example, a VSS_2 line may be inserted above each column of memory cells, as illustrated in FIG. 8. Such an example would result in an insertion frequency similar to that of FIG. 6. However, the configuration of FIGS. 8-9 offers additional flexibility by routing VSS_2 lines above the metal layers of the memory array. This enables a VSS_2 line to be shared between two or more columns of memory cells, since the VSS_2 line can now be widened to extend across multiple columns without affecting memory density.

FIG. 10 illustrates an exemplary top-down view of a system 70 including a memory device (denoted with vertical hatch lines) and one or more subsystems (denoted with dots), all of which are embedded within and/or arranged upon a single semiconductor chip (i.e., a System-on-Chip). In an effort to effectively utilize chip area, many System-on-Chip (SoC) designs divide the memory device into numerous memory blocks. The memory blocks are embedded at various locations within the chip, rather than concentrated in one large memory unit. Substantially any number of memory blocks may be included within system 70; however, some SoC designs may require a relatively

large number of memory blocks (e.g., up to 200 or more) to effectively utilize chip area. As such, the memory blocks should be constructed so as to minimize or completely avoid routing restrictions within the upper-level metallization layer, or "chip-level routing layer", of the system.

FIG. 10 illustrates the embodiment in which VSS_2 lines are formed within the chip-level routing layer, as described above in reference to FIGS. 8-9. Although such an embodiment may limit chip-level signal and power routing to some extent, it allows each of the individual memory blocks to be coupled to a chip-level ground supply grid. As noted above, the ground supply grid can be tailored to eliminate any voltage fluctuations (i.e., voltage droop or ground bounce) that may occur when a relatively large amount of current is discharged onto one of the ground supply lines within the memory array (i.e., a VSS_1 line associated with an asserted row).

Various types of subsystems may be integrated within system 70 including microprocessor and micro-controller cores, digital signal processors (DSPs), communication cores, sound and video cores, radio frequency (RF) cells, power management, and high-speed interfaces, among others. A plurality of transmission lines (not shown) may then be used for interconnecting the subsystems and/or for connecting particular subsystems to one or more memory blocks. In the current embodiment, the plurality of transmission lines (otherwise referred to as chip-level signal and power lines) are routed between the VSS_2 lines within the chip-level routing layer. Various types of transmission lines may be integrated within system 70 including input/output (I/O) lines, clocking lines, intra-system signal lines, and power and ground supply lines.

Several embodiments of an improved memory cell architecture have now been described in the context of a single-port SRAM cell architecture. As noted above, all bitlines are formed in the lowest available metallization layer of the memory array. Since the intrinsic capacitance of a conductive line tends to increase at higher metallization layers, the present improvement minimizes the intrinsic bitline capacitance to improve

memory speed and performance. In addition, all wordlines and at least a portion of the ground supply lines are formed above the bitlines of the memory array. This also enhances memory speed and performance by enabling the wordlines and ground supply lines to vertically shield the bitlines from stray capacitances in overlying transmission lines. A two-dimensional ground supply grid is also provided for reducing the occurrence of voltage droop, ground bounce and electromigration effects in the memory array, thereby improving the robustness of the memory array.

The improvements described above are not limited to a single-port CMOS SRAM architecture, and may be additionally applied to: SRAM cell architectures having more than one port, SRAM cell architectures formed according to different process technologies (e.g., Silicon On Insulator, SOI), other semiconductor memory cell architectures (e.g., DRAM and various non-volatile memories, such as FeRAM and MRAM), and other semiconductor devices (e.g., analog or mixed signal elements, and CMOS based sensor elements, such as temperature, pressure, magnetic and chemical sensors). Additional features and improvements of the present invention will be described below in the context of a dual-port memory cell array.

FIGS. 11-13 illustrate various semiconductor layers and structures that may be used to form a dual-port CMOS SRAM array. Though only a portion of the memory array is illustrated for the sake of clarity, the layout described herein may be extended to form memory arrays of substantially any size.

FIG. 11 illustrates the substrate through first metal layers of a dual-port memory array according to one preferred embodiment of the present invention. More specifically, the substrate through first metal layers of six dual-port SRAM cells -- formed in three rows (denoted R_X, R_{X+1}, and R_{X+2}) and two columns (denoted C_X and C_{X+1}) -- are illustrated in layout 1100 of FIG. 11. Layout 1100 includes the active regions, isolation regions, gate structures and contact structures that may be used to form the NMOS and PMOS transistors of the dual-port memory cell (200) shown in FIG. 2.

The active regions, i.e., the areas where active transistors are to be formed, are embedded within a semiconductor substrate. The semiconductor substrate may be a silicon substrate doped with n-type and p-type impurities in the vicinity of the PMOS and NMOS transistors, respectively. The active regions typically include diffusion regions and isolation regions. Diffusion regions are formed within the active regions adjacent to transistor gate structures and may include, e.g., lightly doped drain regions and heavily doped source/drain regions. Dielectric isolation regions separate active regions from one another, and as such, may include field oxide regions formed by any number of techniques. The diffusion regions and isolation regions may be formed according to any method known in the art.

Each transistor includes a gate structure, which is formed above an active region, arranged between a pair of source/drain regions, and separated from the substrate by a relatively thin dielectric layer. In some cases, the gate structures may be formed from polysilicon (or "poly"), which may be deposited, e.g., by chemical vapor deposition (CVD) of silicon from a silane source, onto the thin dielectric layer overlying the substrate. Other methods of polysilicon formation are known in the art. Gate structures are not limited to polysilicon, however, and may be formed from any suitable conductive material, such as aluminum, titanium nitride, and tantalum nitride, among others. In some cases, the gate structures may include multiple layers of material, such as, e.g., a doped polysilicon and a silicide. For example, a layer of refractory metal (e.g., cobalt, nickel or titanium) may be formed upon a polysilicon layer and heated to induce a reaction between the refractory metal and the polysilicon layer. This reaction may result in the formation of a silicide, such as cobalt silicide, nickel silicide or titanium silicide.

Conductive regions and layers of the memory cell may be isolated from one another by dielectric layers. In addition to the relatively thin dielectric layer mentioned above, a relatively thick dielectric layer (not shown) may be used for isolating the gate structures from an overlying metal layer. Suitable dielectrics may include silicon dioxide (SiO_2), tetraorthosilicate glass (TEOS), silicon nitride (Si_xN_y), silicon oxynitride

($SiO_xN_y(H_z)$), and silicon dioxide/silicon nitride/silicon dioxide (ONO). The dielectrics may be grown or may be deposited by physical deposition such as sputtering or by a variety of chemical deposition methods and chemistries such as chemical vapor deposition. Additionally, the dielectrics may be undoped or may be doped (e.g., with boron, phosphorus, boron and phosphorus, or fluorine) to form a doped dielectric layer such as borophosphosilicate glass (BPSG), phosphosilicate glass (PSG), and fluorinated silicate glass (FSG).

Because the conductive regions and layers of the memory cell are isolated from one another, it is often necessary to form openings in a dielectric layer to provide access to underlying regions or layers. In general, the term "contact opening" or "contact hole" may be used to refer to an opening through a dielectric layer that exposes a diffusion region, or an opening through a dielectric layer arranged between a polysilicon structure and a local interconnect (or a first metal layer). On the other hand, an opening through a dielectric layer arranged between two metal layers may be referred to as a "via". For the purposes of this disclosure, the term "contact opening" will be used to refer to a contact opening and/or a via.

In some cases, contact openings may be filled with a conductive material to form "contact structures." The contact structures provide a pathway through which electrical signals from an overlying conductive region or layer can reach an underlying region or layer of the memory cell. Though any suitable conductive material may be used, metals (such as, e.g., aluminum (Al), copper (Cu) and tungsten (W)) are generally preferred so as to minimize the resitivity of the contact structure. Many types of contact structures (e.g., self-aligned contacts and borderless contacts) may be included within layout 1100. Although square contact structures are illustrated in layout 1100, the contact structures may be formed in any other suitable shape. As described herein, a "suitable" shape may be one that does not increase the complexity of the memory array layout.

FIG. 11A provides a magnified view of the active regions, isolation regions, gate structures and contact structures, which may be used to form the NMOS and PMOS transistors of a dual-port memory cell. As shown in FIG. 11A, the dual-port memory cell (located within row R_{X+1}, column C_X of layout 1100) includes two NMOS active regions and two PMOS active regions. Each of the NMOS active regions comprises a latch transistor and two access transistors. For example, polysilicon segments 1110A, 1110B and 1110C are arranged above N-type diffusion region 1120 to form the gate structures of access transistors T3, T5 and latch transistor $T1_N$. Polysilicon segments 1110A, 1110B′ and 1110C′ are arranged above N-type diffusion region 1130 to form the gate structures of access transistors T4, T6 and latch transistor $T2_N$. In addition, each of the PMOS active regions comprises a latch transistor. For example, polysilicon segments 1110C and 1110C′ also extend across P-type diffusion regions 1140 and 1150 to form the gate structures of latch transistors $T1_P$ and $T2_P$, respectively.

As will be described in more detail below, polysilicon segment 1110A may be coupled to an overlying wordline (e.g., WL_A) through various contact structures and interconnects, and thus, may be referred to herein as the "first local wordline" of the memory cell. As noted above, the term "local" refers to a connection that extends only partially across a memory cell, or stated another way, a connection that does not extend completely from one side of the memory cell to the other. Polysilicon segments 1110B and 1110B′ may also be coupled to an overlying wordline (e.g., WL_B) through various contact structures and interconnects, and thus, may be referred to herein as the "second local wordline" of the memory cell. However, segments 1110B and 1110B′ may be split into distinct portions and arranged on opposite sides of the memory cell.

In one preferred embodiment, each of the first and second local wordlines are shared by two access transistors. In some cases, a local wordline may be shared by two access transistors arranged within the same memory cell. For example, the first local wordline may be shared by access transistors T3 and T4, as shown in FIG. 11A. In other cases, however, a local wordline may be shared by two access transistors arranged within

different memory cells. For example, one portion of the second wordline (i.e., segment 1110B) may be shared between access transistor T5 and an NMOS access transistor within a horizontally adjacent memory cell (located, e.g., within row R_{X+1}, column C_{X-1}), as shown in FIGS. 11 and 11A. The other portion of the second wordline (i.e., segment 1110B') may be shared between access transistor T6 and an NMOS access transistor within another horizontally adjacent memory cell (located, e.g., within row R_{X+1}, column C_{X+1}). In any case, the horizontal dimension, or width, of the memory cell can be reduced by sharing the first and second local wordlines as described herein.

In another preferred embodiment, each transistor of the memory cell shares at least one contact structure with another transistor. In some cases, two or more transistors within different memory cells may utilize a "shared contact structure" for contacting a common semiconductor feature. For example, a contact structure providing access to an overlying bitline (e.g., BL_B) may be shared between an access transistor of the memory cell (e.g., access transistor T5 of FIG. 11A) and an access transistor within a vertically adjacent memory cell located, e.g., within row R_X, column C_X. Similarly, a contact structure providing access to an overlying power supply line (VDD) or ground supply line (VSS) may be shared between a latch transistor of the memory cell (e.g., latch transistor $T1_N$ of FIG. 11A) and a latch transistor within another vertically adjacent memory cell located, e.g., within row R_{X+2}, column C_X. In this manner, a shared contact structure may be used to conserve space within the memory cell.

In conventional memory cell layouts, all elements of a memory cell are usually contained within a "cell pattern boundary." If a contact structure is shared between adjacent memory cells - the contact structure is usually shared at the cell pattern boundary - with substantially half of the contact structure residing on each side of the cell pattern boundary. In other words, contact structures shared between adjacent memory cells are usually symmetrically formed about the boundary between adjacent memory cells.

In contrast, the shared contact structures described herein are preferably offset from the cell pattern boundary. In other words, one or more elements of the memory cell may extend past the cell pattern boundary into an adjacent memory cell. This enables contact structures to be shared unequally between the memory cell and the adjacent memory cell. FIG. 11A illustrates several ways by which shared contact structures may be offset from the cell pattern boundary. In some cases, a shared contact structure may reside fully within the memory cell (*see*, e.g., the VDD, VSS and BLB_B contacts of transistors $T1_P$, $T1_N$ and T6), or may reside fully within the adjacent memory cell (*see*, e.g., the VDD, VSS and BL_B contacts of transistors $T2_P$, $T2_N$ and T5). In other cases, unequal portions of a shared contact structure may be formed within the memory cell and the adjacent memory cell (*see*, e.g., the BL_A and BLB_A contacts of transistors T3 and T4). In any case, the vertical dimension, or length, of the memory cell can be reduced (e.g., about 10% to about 20%) by offsetting the shared contact structures from the cell pattern boundary as described herein. Since bitlines are typically formed along the vertical dimension, a reduction in the vertical dimension reduces the overall bitline length to increase memory speed and memory cell density.

A simple mirroring technique may be used to incorporate the memory cell of FIG. 11A into the memory array layout of FIG. 11. For example, a column of memory cells may be formed by rotating vertically adjacent memory cells about an x-axis and about a y-axis, where the x- and y-axes extend horizontally and vertically, respectively, through a center of each memory cell. Considering that the memory cell of FIG. 11A is located within row R_{X+1}, column C_X of layout 1100, a vertically adjacent memory cell located within either row R_X or row R_{X+2} of column C_X would be formed by rotating a copy of the original memory cell (FIG. 11A) about its x- and y-axes before placing the rotated copy in the vertically adjacent position. In this manner, each memory cell in the column may be considered a "rotated copy" of a memory cell located directly above and/or below the rotated copy. If the memory array includes more than one column of memory cells, the original column may be reproduced at a horizontally adjacent location, thereby forming multiple rows of memory cells. As such, each memory cell in the row may be

considered an "exact copy" of a memory cell located directly alongside the exact copy. The memory array layout of FIG. 11 is assembled in the design environment, so that rotation and mirroring of the individual memory cells may be performed within a software application. For a given layer, the memory array may be printed onto a reticle (along with other subsystem circuits on the chip) for subsequent transfer to a semiconductor wafer.

The above-mentioned mirroring technique enables additional features and advantages to be incorporated into layout 1100 of FIG. 11. In one embodiment, the active regions of layout 1100 are formed substantially parallel to one another; the gate structures of layout 1100 are also formed substantially parallel to one another, though perpendicular to the active regions. Since all transistors are formed in substantially the same direction, any systematic differences that may exist between perpendicularly formed transistors are eliminated by the present embodiment.

Forming all transistors in the same direction also eliminates the need for active regions that are formed perpendicular to one another and/or formed in an "L-shape." In a preferred embodiment, the N-type diffusion regions of layout 1100 are each formed as a substantially continuous line of constant width, where a "continuous line" is described as one that extends from one side of the memory array to the other. Thus, two N-type diffusion regions may be formed within each column of memory cells and shared by all NMOS transistors within that column. Though the P-type diffusion regions of layout 1100 are each formed as a substantially straight line, each P-type diffusion region extends only partially across two vertically adjacent memory cells. Thus, each P-type diffusion region may be shared by two PMOS transistors, one residing within each of the vertically adjacent memory cells. Therefore, the present embodiment may further reduce the width and length of the memory cell by avoiding complex geometries in the active regions and sharing diffusion regions between two or more vertically adjacent cells. This has the advantage of simplifying the photolithography process and reducing the memory cell density.

In another preferred embodiment, a rectangular-shaped isolation region is formed within each N-type diffusion region of layout 1100 for terminating access transistors T3-T6. For example, a distal end of the first local wordline (e.g., segment 1110A) and a distal end of one portion of the second local wordline (e.g., segment 1110B') may be terminated over one of the rectangular-shaped isolation regions. However, the first and second local wordlines are preferably formed such that their distal ends are horizontally and vertically spaced from one another. In some cases, a rectangular-shaped isolation region may be shared between two vertically adjacent memory cells, as shown in FIG. 11. Thus, the present embodiment further aids in reducing memory cell density by terminating two or more local wordlines in a staggered formation above the isolation region. This has the advantage of reducing memory cell density by reducing the width and length of the memory cell.

FIG. 12 illustrates the first through second metal layers of the dual-port memory array according to one preferred embodiment of the present invention. In general, layout 1200 includes a first metal layer and a second metal layer, which is dielectrically spaced above the first metal layer shown in phantom in layout 1100 of FIG. 11.

In some embodiments, the first metal layer may be used as a local interconnect layer for cross-coupling internal nodes of the NMOS and PMOS transistors used to form the SRAM array. The local interconnect layer may also be used for coupling overlying conductive layers to the underlying transistors. Note, however, that reference to the local interconnect layer as a "metal layer" does not limit the constituents of that layer to only metallic materials (e.g., Al and Cu). Instead, local interconnects may be fabricated from any conductive material known in the art, such as, e.g., polysilicon, doped polysilicon, refractory metal (e.g., W), silicide, or a combination of these materials.

After forming a dielectric layer (not shown) upon the first metal layer, one or more contact openings (labeled "Via1" in FIG. 12) may be etched into the dielectric layer to provide access to the first metal layer. A conductive layer may then be formed above

the dielectric layer by depositing a metallic material onto the surface of the dielectric layer. After the conductive layer is patterned and etched, the conductive layer may be referred to as a "second metal layer."

In a preferred embodiment, the bitlines of the memory array are formed within the second metal layer. As mentioned above, forming all bitlines within the second metal layer (or lower) may advantageously reduce the intrinsic capacitance of the bitlines to attain faster memory addressing times. If the memory array comprises more than one port, horizontal capacitive shielding may be provided within the second metal layer between bitlines of dissimilar port.

Capacitive isolation between bitline ports may be especially important in dual-port memory arrays, which allow simultaneous read/write access to cells within the same column via Port A and Port B bitlines. For example, Port A bitlines may be used to perform a read operation on a memory cell, while Port B bitlines are simultaneously used to perform a write operation on another memory cell within the same column. Since a maximum voltage range is applied between the bitlines during the write operation, the write operation on the Port B bitlines may induce a significant charge through capacitive coupling on the Port A bitlines. Such cross-coupling may significantly slow down the read operation and/or cause errors to occur on the Port A bitlines. A similar event may occur when Port A and Port B bitlines are simultaneously used to perform separate read operations on two different memory cells within the same column; the mutual capacitive cross-coupling may slow the read operation within both ports.

In a preferred embodiment, ground supply lines (VSS_2) may be formed between and substantially parallel to the Port A and Port B bitlines (i.e., between BL_A and BL_B, and between BLB_A and BLB_B) of a multiple-port memory array to prevent inter-port capacitive coupling. Such a case is illustrated in the dual-port memory array of FIG. 12. In other cases, power supply lines (VDD) may be used to provide capacitive shielding between bitlines of dissimilar port. Thus, the present embodiment increases memory

performance by inserting VSS (or VDD) lines between the Port A and Port B bitlines. Though the VSS (or VDD) lines may slightly increase the overall bitline capacitance, their inclusion between the Port A and Port B bitlines improves memory performance by significantly reducing capacitive coupling therebetween.

FIG. 13 illustrates the second through third metal layers of the dual-port memory array according to one preferred embodiment of the present invention. In general, layout 1300 includes second metal layer and a third metal layer, which is dielectrically spaced above the second metal layer shown in layout 1200 of FIG. 12. The third metal layer of the memory array may be formed in a manner similar to the formation of the second metal layer. For example, a dielectric layer (not shown) may be formed upon the second metal layer, and one or more contact openings (labeled "Via2" in FIG. 13) may be etched into the dielectric layer for accessing the second metal layer. After a metallic material is deposited onto the surface of the dielectric layer, the metallic material may be patterned and etched to form the "third metal layer".

Though non-metallic conductive materials (e.g., silicide and polysilicon) may be used to form the second and third metal layers, a metallic material is generally preferred due to the lower resitivity of metallic (e.g., about 0.2 ohms/square-unit) versus non-metallic conductive materials (e.g., about 20 to 50 ohms/square-unit). Examples of suitable metals include aluminum (Al), copper (Cu), Silver (Ag) and gold (Au). Because of their lower resitivity, metal conductors can be much longer than those of polysilicon or silicide. For this reason, metal conductors within the memory array (e.g., the bitlines, wordlines, power and ground supply lines) may extend across the entire memory array, or at least a portion thereof when the memory array is broken into numerous memory blocks.

In a preferred embodiment, the wordlines of the memory array are formed within the third metal layer. By forming wordlines above the bitlines, the wordlines function to vertically shield the bitlines from any transmission lines that may be routed over the memory array for transporting chip-level signals. The vertical shielding provided by the

wordlines minimizes the adverse affects of stray capacitances from the transmission lines, thereby protecting bitline signal development during read operations and reducing the occurrence of data corruption in the "sensed" signal.

If the memory array comprises more than one port, horizontal capacitive shielding may also be provided within the third metal layer between wordlines of dissimilar port. For example, unnecessary voltage spikes may occur in at least one wordline of a dual-port memory array when the wordlines of both ports are asserted concurrently. This may cause increased leakage and/or data corruption in one or more memory cells of the array. Since the ports of a dual-port memory cell are independently operated, a situation may occur in which one of the wordlines (e.g., Port A wordline) is ramping up in voltage, while the other wordline (e.g., Port B wordline) is ramping down in voltage. In this situation, any significant capacitive coupling between the Port A and Port B wordlines can lead to a delay in "turning off" the Port B wordline and/or a delay in the WL-to-BL separation time.

In a preferred embodiment, ground supply lines (VSS_1) may be formed between and substantially parallel to the Port A and Port B wordlines (i.e., between WL_A and WL_B) of a multiple-port memory array to prevent inter-port capacitive coupling. Such a case is illustrated in the dual-port memory array of FIG. 13. In other cases, power supply lines (VDD) may be used to provide capacitive shielding between wordlines of dissimilar port. Thus, the present embodiment increases memory speed and performance by inserting VSS (or VDD) lines between the Port A and Port B wordlines.

Because the VSS_1 lines are perpendicular to the bitlines of the memory array, however, a substantially large amount of current may be discharged onto a single VSS_1 line during a read operation. To accommodate this potentially large discharge current, the ground supply lines (VSS_1) within the third metal layer may be coupled to the ground supply lines (VSS_2) within the second metal layer to form a two-dimensional power supply grid. In doing so, the adverse effects of voltage droop, ground bounce and

electromigration can be reduced, or even avoided, by interconnecting the VSS_1 and VSS_2 lines at an appropriate frequency. In some cases, the VSS_1 and VSS_2 lines may be coupled within each cell of the memory array. However, it may only be necessary to couple the VSS_1 and VSS_2 lines once every X-number of rows and Y-number of columns (e.g., at every row and every 8 to 32 columns), where X and Y are determined by the restrictions set for avoiding voltage droop, ground bounce and electromigration.

In one embodiment, a power supply line may be included within the memory array for every two rows of memory cells, as shown in FIG. 13. In some cases, the power supply line (VDD) line may be shared between two vertically adjacent rows by forming the VDD along the cell pattern boundary. This may further increase memory density by reducing the vertical dimension, or length, of the memory array.

It will be appreciated to those skilled in the art having the benefit of this disclosure that this invention is believed to provide an improved memory architecture offering substantial increases in memory density, speed and performance, in addition to reduced congestion in upper-level metallization layers of a system. Further modifications and alternative embodiments of various aspects of the invention will be apparent to those skilled in the art in view of this description. It is intended that the following claims be interpreted to embrace all such modifications and changes and, accordingly, the specification and drawings are to be regarded in an illustrative rather than a restrictive sense.